Titel: Verfahren zur Erstellung nutzer- und situationsspezifischer, generativer Inhalte

Titel (englisch): Method for generation of user- and situation-specific generative content

Autoren: Svenja Kobel, Dr. Christian Wiede, Dr. Roman Seidel

Institutsadresse: Fraunhofer-Institut für Mikroelektronische Schaltungen und Systeme IMS, Finkenstr. 61, 47057 Duisburg

Kurzfassung (deutsch):

Die vorgeschlagene Lösung beschreibt ein neuartiges, multimodales Dialogsystem, das LLMs mit Emotionserkennung und kontaktloser Vitalparametermessung in einem Kiosk/Terminal verbindet. Im Unterschied zum Stand der Technik werden Sprache (Text, Audio), Videodaten, rPPG-basierte Herz- und Atemfrequenz sowie eine RAG-gestützte Wissensdatenbank in einer fusionierten Zustandsanalyse vereint. Ein Modul zur emotionssensitiven Dialogsteuerung passt Systemprompts, Tonfall, Gesprächsdauer und Modalität der Antwort dynamisch an den emotional-physiologischen Zustand an. Die Ausgabe erfolgt über TTS, Stereo-Lautsprecher, einen 3D-Avatar mit Emotion-Spiegelung, Visualisierung von Vital- und Umgebungsdaten sowie Schnittstellen zu externen Systemen (z. B. Heizungssteuerung).

Kurzfassung (englisch):

The proposed solution describes a novel, multimodal dialogue system that combines LLMs with emotion recognition and contactless vital parameter measurement in a kiosk/terminal. Unlike the state of the art, speech (text, audio), video data, rPPG-based heart and respiratory rates, as well as a RAG-supported knowledge database are integrated into a fused state analysis. A module for emotion-sensitive dialogue control dynamically adjusts system prompts, tone, conversation duration, and response modality according to the emotional-physiological state. The output is delivered via TTS, stereo speakers, a 3D avatar with emotion mirroring, visualization of vital and environmental data, as well as interfaces to external systems (e.g., heating control).

Fünf Stichwörter (deutsch): Generative KI, Agentische KI, Vitalparameter, Terminal, Emotionsanalyse

Fünf Stichwörter (englisch): Generative AI, Agentic AI, Vital Parameters, Terminal, Emotion Analysis

Einleitung und Motivation

Generative KI erlebt derzeit einen Hype. Modelle bzw. Chat Bots wie GPT5, Gemini oder Claude werden von einer großen Masse an Leuten weltweit genutzt. Diese Modelle beruhen auf den sogenannten großen Sprachmodellen (kurz LLM für Large Language Model in Englisch). Sie berechnen anhand einer Eingabe, auch Prompt genannt, eine Ausgabe, welche der statistischen Wahrscheinlichkeit entspricht. Die Modelle sind dabei in der Lage, sich auf die jeweilige Wortwahl der Prompts einzugehen und dabei auf den Nutzer einzugehen. Während dies textuell passiert, wird außer Acht gelassen, dass 90 % der Interaktion zwischen Menschen nonverbal erfolgt. Durch Gestik, Mimik und Betonung der Sprache werden ebenfalls Informationen transportiert, welche für unsere tägliche Kommunikation unerlässlich sind. Diese Informationen fehlen den heutigen Chat Bots bisher, sodass ihrer interaktiven, natürlichen Nutzung noch Einschränkungen unterliegt.

Stand der Technik

Emotionsanalyse bzw. Sentimentanalyse zur Personalisierung der User Experience in digitalen Services wie Chatbots für eine intelligente, konversationelle Steuerung ist weit verbreitet. Neben der Analyse von Texteingaben (Wortwahl, Satzbau etc.) kommen visuelle und auditive Verfahren zum Einsatz, etwa Facial Emotion Recognition über Kamerabilder, Speech Emotion Recognition über Stimmmerkmale sowie teilweise thermische Sensorik [1 - 8]. In diesen Ansätzen werden jedoch keine Vitalparameter erfasst, um auch Stress, Erregungsniveau und kognitive Belastung analysieren zu können.

Einige Ansätze gehen darüber hinaus und integrieren zusätzlich kontaktbasierte Vitalparametermessung mithilfe von Wearables, um Eingaben besser zu interpretieren und nuancierte, empathische sowie handlungsleitende Reaktionen zu generieren [9 - 11]. Diese Lösungen setzen jedoch das Tragen der Sensorik voraus, was in der Praxis zu Nachteilen führt: Geräte werden vergessen oder nicht regelmäßig genutzt, es besteht ein Hygienerisiko und es können Hautirritationen oder Trageunbehagen (Discomfort) auftreten.

Auch andere kontaktbasierte Sensorik wird eingesetzt. So existieren Beispiele mit integriertem PPG-Sensor in haptischen Gegenständen (z. B. Stofftier) zur Vitalparametermessung [12]. In Kombination mit weiteren Sensoren entstehen so interaktive Begleiter mit Emotionserkennung und affektiver Sprachinteraktion, etwa für ältere Menschen. Auch ein Handschuh-System wurde vorgestellt [13]. Die genannten Nachteile kontaktbasierter Systeme gelten jedoch auch hier. Hervorzuheben ist das Hygienerisiko bis hin zur Infektionsgefahr, insbesondere wenn Produkte von mehreren Personen verwendet werden.

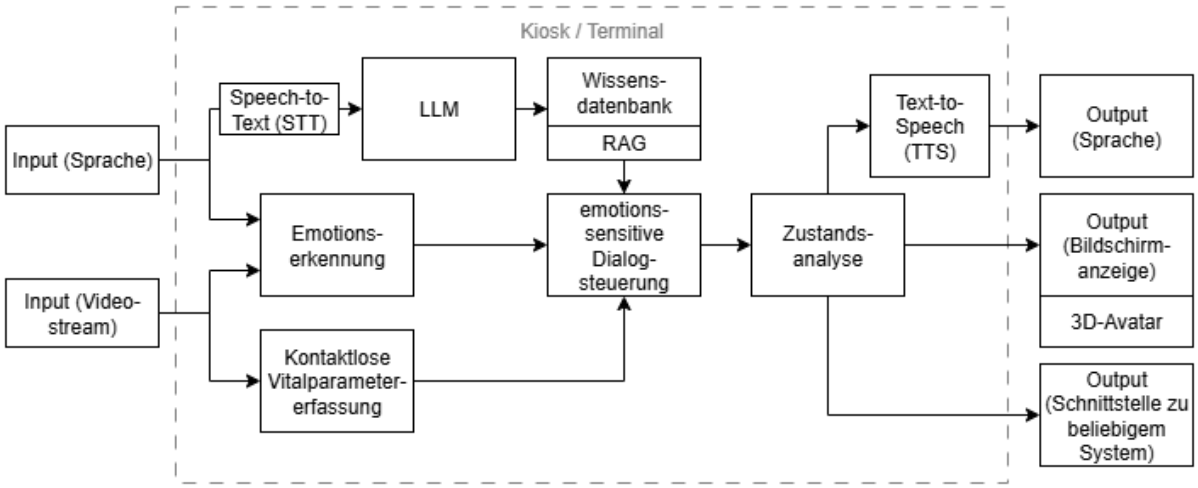
Zudem existieren Systeme, die Emotionserkennung mit kontaktloser Biosignalerfassung verbinden, beispielsweise das Patent „EDGE-AI POWERED MULTIMODAL EMOTION RECOGNITION SYSTEM WITH CONTACTLESS PHYSIOLOGICAL SENSING AND ADAPTIVE FUSION“ [14]. Dabei handelt es sich jedoch um ein rein multimodales Erkennungsmodell mit kontaktloser Biosignalerfassung, also ohne Dialog, ohne LLM und ohne RAG-Komponenten, sodass keine adaptive Gesprächssteuerung auf Basis der Zustandsanalyse erfolgt.

Ein Patent zu einem „SMART MIRROR WITH HEALTH DIAGNOSTICS“ [15] geht einen Schritt weiter und beschreibt ein smartes Spiegel-Gesundheitsdiagnostiksystem für berührungslose, ganzheitliche Health-Screenings mittels Bildgebung, KI- und Telemedizin-Funktionen. Dennoch findet auch hier keine biosignalbasierte, adaptive Dialogsteuerung statt. Vielmehr arbeiten Biosignale und LLM parallel zur Gesundheitsdetektion zusammen, wobei das LLM die Beantwortung gesundheitsbezogener Nutzerfragen ermöglicht, ohne dabei Parameter wie Gesprächslänge, Timing oder Tonfall biosignalbasiert und emotionssensitiv an den physiologischen Zustand anzupassen.

In dieselbe Richtung weist die Arbeit von Kasno et al. [16], die Facial Expression Analysis, Remote Photoplethysmography (rPPG) und Conversational AI kombiniert. Auch hier dient die KI der therapeutischen Gesprächsführung auf Basis gemessener Daten. Eine adaptive Gesprächssteuerung auf Basis fusionierter Zustandsindikatoren erfolgt jedoch ebenfalls nicht, sodass eine echtzeitfähige Anpassung von Gesprächsparametern ausbleibt.

Zusammenfassend ist kein System bekannt, das Sprache, Audiodaten, Videodaten, LLM, Emotionserkennung und kontaktlose Vitalparametermessung in einer fusionierten Zustandsanalyse so vereint, dass eine emotionssensitive, biosignaladaptive Dialogsteuerung möglich wird und zusätzlich eine Wissensdatenbank inklusive RAG-System integriert ist, um externes Wissen dynamisch im System verfügbar zu machen.

Methodik



Input (Sprache): Ein Mikrofon erfasst die natürliche Sprache der Person, um sie an das Speech-to-Text (SST) Modul weiterzugeben. Das Mikrofon ist in Richtung der sprechenden Person orientiert und verfügt über Rauschunterdrückung, um Umgebungsgeräusche zu reduzieren. Das Mikrofon muss ausreichend hochaufgelöst sein, um den „Tonfall" (Sprachstimmung) klar aufzunehmen und sollte über eine Nierencharakteristik verfügen und einen niedrigen Klirrfaktor haben, um die Aufnahme in halligen Räumen zu ermöglichen.

Input (Videostream): Der Input für die Emotionserkennung und die kontaktlose Vitalparametererfassung ist ein Videostream von einer Kamera. Die Kamera ist als RGB-Kamera vorgesehen und erfasst den Kopf und den Torso der Person. Dabei sollte die Person frontal oder leicht schräg von der Kamera erfasst werden. Eine externe Beleuchtung ist zusätzlich möglich. Eine Realisierung mit normalem Tageslicht funktioniert ebenfalls.

Speech-to-Text (STT) und LLM: Natürliche Sprache wird über ein Speech-to-Text Modul in Text umgewandelt. Hierfür kommen Transkribierungsverfahren zum Einsatz, wie beispielsweise Whisper [17] oder LiveKit [18], die in Echtzeit auf lokaler Hardware mit beschränkten Rechenressourcen Sprache in Text konvertieren. Im Anschluss wird der erzeugte Text (z.B. eine gestellte Frage) von einem Large Language Model verarbeitet. Das LLM basiert auf einer Transformer-Architektur, um Muster, Bedeutungen und Zusammenhänge in Sprache zu erkennen. Basierend auf den gelernten Trainingsdaten gibt das Modell diejenige Antwort aus, die mit hoher Wahrscheinlichkeit auf die gestellte Frage passt.

Emotionserkennung: Die Emotionserkennung basiert auf dem Sprachinput und dem Videostream. Der Sprachinput wird auf Textmerkmale (z. B. Wortwahl, Satzbau, semantischer Kontext) sowie auf stimmliche Merkmale (z. B. Tonhöhe, Lautstärke, Tempo, Pausen, Stimmfarbe) geprüft. Die Videodaten ermöglichen die Erkennung von Gesichtsemotionen mittels visueller Merkmale (z. B.

Gesichtslokalisierung, Gesichtslandmarken, Mimikdynamik wie Augenbrauen- und Mundbewegungen, Kopfhaltung). Die Ergebnisse aus den Modalitäten Text, Audio und Bild werden fusioniert (z. B. über gewichtete Modelle) und als Emotionsschätzung mit Konfidenzwerten ausgegeben.

Kontaktlose Vitalparametererfassung: Die kontaktlose Vitalparametermessung basiert auf dem Videostream. Dabei wird die Herzfrequenz mittels remote Photoplethysmographie (rPPG) gemessen. rPPG detektiert Veränderungen der Reflexion von Umgebungslicht auf der Hautoberfläche, die durch Blutvolumenschwankungen während jedes Herzzyklus verursacht werden. Hierzu werden im Videostream der Gesichtsbereich und darin enthaltene relevante Regionen von Interesse (etwa Stirn und Nase) bestimmt und über die Zeit stabil verfolgt. Zur Erhöhung der Robustheit wird die Erkennung des Gesichts regelmäßig überprüft und bei Bedarf aktualisiert. Aus diesen Regionen werden pro Bild die mittleren Farbwerte (Rot, Grün, Blau) zu Zeitreihen zusammengeführt, anschließend durch geeignete Filter von Störungen bereinigt und mittels Kanalkombination bzw. Transformation zu einem rPPG-Signal kombiniert. Die Herzfrequenz wird fortlaufend durch spektrale Analyse im physiologischen Frequenzbereich ermittelt. Die Atemfrequenz kann aus Modulationen des Herzfrequenzsignals oder durch Analyse der Brustkorbbewegung erfasst werden. Für die Analyse der Brustkorbbewegung wird im Videostream relativ zur erkannten Gesichtsregion der Brustbereich lokalisiert. In diesem Bereich werden markante Bildpunkte detektiert und über die Zeit verfolgt. Aus den Bewegungsverläufen wird eine repräsentative Trajektorie der Auf- und Abbewegung des Brustkorbs gebildet und zu einem respiratorischen Signal zusammengeführt. Das gewonnene Signal wird vorverarbeitet, um Störungen zu reduzieren, langsame Drifts zu entfernen und auf den relevanten Frequenzbereich einzuschränken. Die Atemfrequenz wird anschließend durch spektrale Analyse bestimmt.

Wissensdatenbank und RAG: Um externes Wissen im System verfügbar zu machen ist eine Wissensdatenbank und die Abfrage mittels RAG-System essenziell. Die Wissensdatenbank enthält unterschiedliche Dokumente, die für die Interaktion des Kiosks / des Terminals mit dem Endnutzer notwendig sind. Dies können zum Beispiel Handbücher, Listen, Datenbanken, E-Mails, Webseiten oder auch Metadaten sein. Das RAG-System besteht aus einem Embedding Modell, das die Textdaten vektorisiert und in einer Vektordatenbank ablegt. Dafür müssen zuvor die Dokumente sinnvoll in thematisch zusammenhängende Abschnitte eingeteilt werden („chunking"), wobei die chunking-Größe je nach Anwendungsfall und Dokumententyp dynamisch gewählt wird. Mithilfe eines Wissensgraphs werden Zusammenhänge zwischen den Daten und den Metadaten modelliert, die es dem Endnutzer ermöglichen, z.B. zeitliche Verläufe (Buchungen, Zeitreihen) abzufragen.

Emotionssensitive Dialogsteuerung: Dieses Steuerungsmodul speist sich aus hinterlegtem Wissen, der Emotionserkennung und aus den Ergebnissen der kontaktlosen Vitalparametererfassung. Diese Informationen werden fusioniert und so aufbereitet, dass das System je nach „Gemütszustand" der interagierenden Person antwortet. Beispiele für Gemütszustände können aufgeregt (hohe Herzrate, hoch Sprechfrequenz) oder entspannt (niedrige Respirationsrate, langsames Sprechen) sein. Je nachdem, wie dieser Zustand bewertet wird, wird die Wahl des Systemprompts für das lokale LLM gewählt, sodass sich die Audio-Ausgabe des Systems an die des Inputs anpasst.

Zustandsanalyse: Dieses Modul analysiert den Zustand der Person, die mit dem Kiosk / dem Terminal interagiert. Das Modul hat unterschiedliche Lokale LLMs mit Informationen (RAG-System) hinterlegt und passt den Systemprompt dynamisch je nach Ausgangszustand an. Die Analyse des Zustands wird von der Dialogsteuerung entgegengenommen. Zusätzlich entscheidet das Modul, ob ein Output in Form von Sprache, eine Visualisierung auf dem Bildschirm und/oder die Weiterverarbeitung in einem per Schnittstelle angebundenen System erfolgt.

Text-to-Speech(TTS): Soll das System eine Sprachausgabe vollziehen, wird der hinterlegte Antworttext von Text in eine natürlichsprachige Audio-Ausgabe umgewandelt. Dies erfolgt z.B. mit dem Whisper-Modell [17], das lokal im Terminal abgelegt ist. Es wird auf eine hohe Performance und geringe Latenz (Eingabe zu Ausgabe) geachtet.

Output (Sprache): Die Ausgabe von Sprache über ein Terminal erfolgt mithilfe von Stereo-Lautsprechern, die in dem Kiosk / dem Terminal verbaut sind.

Output (Bildschirmanalyse): Auf dem Bildschirm wird ein 3D Avatar angezeigt, der die Lippenbewegungen der Sprache ausführt und die Emotionen der interagierenden Person spiegelt. Zusätzlich wird Temperatur, Luftfeuchte, Vitalparameter, Auslastung des Gebäudes (Büro/Wohnen) und personalisierte hinterlegte Parameter angezeigt, sobald die Person identifiziert wird. Ebenfalls angezeigt wird eine Karte des Gebäudes zur Orientierung und die gesprochene Audio-Ausgabe als Text.

Output (Schnittstelle zu beliebigem System): Durch eine API-Schnittstelle lässt sich jedes beliebige System anbinden. Ein Bespiel ist die Heizungssteuerung, sodass das System die Gewohnheiten des Co-workers oder Bewohnenden hinsichtlich Raumtemperatur automatisch berücksichtigt, sobald sich die Person am Kiosk / am Terminal identifiziert.

Quellen:

[1] Pereira, P., Moniz, H. and Carvalho, J.P. Deep emotion recognition in textual conversatio Rev 58, 10 (2025). https://doi.org/10.1007/s10462-024-11010-y.

[2] Kovacevic N., Holz C., Gross M. and Wampfler R. On Multimodal Emotion Recognition fo Interaction in the Wild. In Proceedings of the 26th International Conference on Multimodal Intera Association for Computing Machinery, New York, NY, USA, 12–21 (2024). https://doi.org/10.114

[3] Iyer A. A., Vojjala S. and A. J. Augmenting Sentiments into Chat-GPT Using FacialEmotio International Conference on Advanced Computing and Communication Systems (ICACCS), Coim (2024). doi: 10.1109/ICACCS60874.2024.10717316.

[4] Tapala, J. Emotion analysis in the development of chatbot user experience. Haaga-Helia U Sciences. Master Thesis (2024).

[5] BERNARD ERIK HOWARD: ARTIFICIAL INTELLIGENCE (AI) CHATBOT SYSTEM WITH AVAT ADVANCED EMOTIONAL CAPABILITIES. US2025293997 AA, 18.09.2025.

[6] CALASCIONE ANTHONY ROBERT: PERSONALIZED ARTIFICIAL INTELLIGENCE DRIVEN MA US2025166007 AA, 22.05.2025.

[7] SR UNIV: AN AI-POWERED COMMUNICATION IDOL SYSTEM FOR REDUCING DIGITAL AI IN202541018658 A, 03.03.2025.

[8] SENTIANTAL PVT LTD: AI-DRIVEN DEVELOPER TRAINING AND INTERVIEW SIMULATION S A, 03.07.2025.

[9] Yu, L., Ge, Y., Ansari, S., Imran, M. and Ahmad, W. Multimodal Sensing-Enabled Large L Automated Emotional Regulation: A Review of Current Technologies, Opportunities, and Challen 4763 (2025). https://doi.org/10.3390/s25154763.

[10] SHENZHEN XIXING INTELLIGENT TECHNOLOGY CO LTD: ADAPTIVE INTERACTIVE LANGU ENABLED BY MULTIMODAL AFFECTIVE COMPUTING AND ITS SUPPORTING METHOD. CN12061

[11] SAMSUNG ELECTRONICS CO LTD: METHOD AND SYSTEM FOR GENERATING PERSONAL IN202311056014 A, 28.02.2025.

[12] Schnitzer, B., Vural, U. C., Schnitzer, B., Sardar, M., Fuerst, O. and Korn, O. Prototyping Robot Companion with Emotion Recognition and Affective Voice Interaction for Elderly People. P on Human-Computer Interaction. 8. 1-32 (2024). doi: 10.1145/3660244.

[13] Vijayalakshmi S., Disney Sandhya A. and Nimala Deve P. AI-Powered Holistic Mental Heal Facial Emotion Recognition, Chatbot, and Voicebot for Personalized Support. International Confe Agents & Artificial Intelligence (ICDSAAI), Chennai, India, pp. 1-6 (2025). doi: 10.1109/ICDSAAI6

[14] SR UNIV: EDGE-AI POWERED MULTIMODAL EMOTION RECOGNITION SYSTEM WITH CO PHYSIOLOGICAL SENSING AND ADAPTIVE FUSION. IN202541021126 A, 21.03.2025.

[15] NIMS UNIV RAJASTHAN JAIPUR: SMART MIRROR WITH HEALTH DIAGNOSTICS. IN20251

[16] Kasno M. A. and Jung J. W. Feasibility of an AI-Enabled Smart Mirror Integrating MA-rPP Conversational Guidance in Realtime. Sensors (Basel). 25(18):5831 (2025). doi: 10.3390/s251858

[17] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). recognition via large-scale weak supervision. In *International conference on machine learning* (pp

[18] https://github.com/livekit/agents, zuletzt abgerufen am 16.10.2025